Project Summary: Mean Field Asymptotics in Statistical Inference: Variational Approach, Multiple Testing, and Predictive Inference

Overview

The new era of big data poses unprecedented statistical and computational challenges in high-dimensional statistical inference. One particular challenge is the "dual objective" nature of various statistical inference tasks: statisticians hope to design procedures that achieve near-optimal statistical efficiency and satisfy desired validity guarantees even under model misspecification. For example, in the task of false discovery rate (FDR) control, statisticians would like to design procedures that incorporate prior knowledge to make more discoveries while controlling FDR even when prior knowledge is inaccurate. Unfortunately, these two objectives compete with each other and are often difficult to be achieved simultaneously.

Furthermore, among statistical inference procedures involving a Bayesian component, performing exact Bayesian inference on large-scale datasets is computationally challenging. Statisticians often use variational inference as a computationally tractable approximation for Bayesian inference. Although a variety of variational inference methods were developed in the past few years, the theoretical guarantees of many variational methods are not yet well-established, even in the well-specified statistical models with planted signals.

The PI will address the above statistical and computational challenges in a few high dimensional statistical inference tasks. As a summary, the proposed project has two complementary goals: 1) design statistically efficient procedures with desired validity guarantees even under model misspecification; 2) develop efficient variational inference algorithms with theoretical guarantees.

Intellectual merit

Focusing on a few stylized problems and backed by extensive preliminary results, the proposed program consists of three major research thrusts: 1) analyze the non-convex landscape of TAP variational inference objective functions and design efficient algorithms for optimizing these functions; 2) in the task of FDR control, design procedures that maximize the number of discoveries when models are correctly specified while controlling the frequentist FDR even under model misspecification; 3) in the task of predictive inference, design procedures that give reasonably small prediction sets while maintaining the frequentist validity of coverage in the presence of model misspecification. This research will develop new techniques for studying the mean field asymptotics of high-dimensional statistical models, which will likely be applicable beyond the specific statistical models and will be relevant in other areas of science and engineering.

Broader impacts

The theoretical results and statistical procedures developed in this project will have a direct impact on various science and engineering applications such as genomics, clinical trials, computer vision, functional magnetic resonance imaging, etc. The developed theoretical tools will be of great interest to other subjects including probability theory and theoretical computer science. The project is also tightly integrated with an education, training and outreach plan. The PI plans to further develop the advanced-level course of "Mean Field Asymptotics in Statistical Learning" to incorporate the proposed research findings, supervise graduate students to work on the proposed research program, and organize workshops and invited sessions on high-dimensional statistical inference at UC Berkeley and various other venues. The PIs will involve female and underrepresented groups through long-term mentorships and diversity-promoting activities, actively recruit and train students with diverse backgrounds for the proposed research program.